Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE® GREB–08-01

Potential Impact of Context Effects on the Scoring and Equating of the Multistage GRE® Revised General Test

Tim Davey

Yi-Hsuan Lee

June 2011

Potential Impact of Context Effects on the Scoring and Equating of the Multistage GRE® Revised General Test

Tim Davey and Yi-Hsuan Lee

ETS, Princeton, NJ

GRE Board Research Report No. GREB-08-01

ETS RR-11-26

June 2011

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do no necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Educational Testing Service
Princeton, NJ 08541

Abstract

Both theoretical and practical considerations have led the revision of the Graduate Record Examinations® (GRE®) revised General Test, here called the rGRE, to adopt a multistage adaptive design that will be continuously or nearly continuously administered and that can provide immediate score reporting. These circumstances sharply constrain the available options for equating or linking together the thousands of unique test forms that will need to be administered each year. The only practical method is to pre-equate each form through item response theory (IRT) methods. Unfortunately, doing so means that item position and context effects are both vitally important and potentially difficult to control. This study examines item position effects in a GRE context and attempts to assess the extent to which they may affect the rGRE. It was found that position effects are in fact present in GRE data and that they can be particularly troublesome under multistage testing. Pretest strategies are able to mitigate effects to some degree, but complete control will require careful attention be paid to the time limits and associated test configuration details in the design of the rGRE.

Key words: context effects, position effects, randomization

Table of Contents

List of Tables

Page

List of Figures

Page

1. Background

Item performance is generally considered to be fragile and subject to change due to a variety of factors and influences (Brennan, 1992). Performance is known to be sensitive to minor changes in wording, format, or the specific way in which an item is presented (Beaton & Zwick, 1990; Burke, Hartke, & Shadow, 1989). The position of an item in a test can also strongly impact its difficulty, while more subtle effects may be induced by the complex relationships between an item and the other items that surround it on a test form (Dorans & Lawrence, 1990; Haladyna, 1992; Harris, 1991).

Most generally defined, *context* refers to the collection of factors noted above, including the position in a test in which an item appears as well as the content, format, and specific features of the other items that surround it.

A number of operational practices have been devised for dealing with context effects and the possibility of changed item performance. The simplest and most important of these practices is to fix the specific wording and presentation of an item throughout its life cycle. This means that an item is not changed between the time it is pretested—and its performance characteristics established—and the time it is used as a scored item on a test form. Should changes be required, the revised version of the item would again undergo pretesting, essentially treating it as a new item with unknown characteristics.

Changes in item performance across time and use complicate maintenance of a testing program in at least two ways. The first, and less serious, concerns the degree to which each of a series of alternate test forms are equally difficult and reliable. High-stakes testing programs assemble forms to exacting specifications but depend on items performing as their pretest statistics predict. If item performance changes dramatically, new forms may in fact be easier or more difficult than desired. Although equating is intended to correct for such differences, equating works best when it's needed least.

A second, much more serious impact of item performance change is on the equating process itself. Common-item equating assumes that the anchor items that link a pair of test forms perform identically. Observed differences in performance on these items can then be solely attributed to differences in the examinee samples assigned each form. Any change in the items themselves confounds the equating process and so degrades the comparability of scores across alternate test forms. Again, a number of operational practices have been developed to minimize

the likelihood of item change. For example, anchor items are generally required to appear in comparable positions across the forms being linked (Kolen & Brennan, 1995). This practice limits position effects, which are well known to affect item difficulty. Anchor items may also be segregated within their own test section to control any interactions with the other items on each of the forms being linked.

2. Context and the New Multistage Graduate Record Examinations Revised General Test (rGRE)

Both theoretical and practical considerations have led to adopting a multistage adaptive design for the revision of the Graduate Record Examinations® (GRE®), here called the Graduate Record Examinations revised General Test or rGRE, which will be continuously or nearly continuously administered and which can provide immediate score reporting. These circumstances sharply constrain the available options for equating or linking together the thousands of unique test forms that will need to be administered each year. Such a large number of forms are needed both to provide adaptation across examinees and to limit the security risks inherent in the re-use of intact forms across occasions. The only practical method for dealing with these circumstances is to pre-equate each form through item response theory (IRT) methods. Unfortunately, doing so means that context effects are both vitally important and potentially difficult to control.

Although the theory behind IRT is silent on (or, more accurately, dismissive of) the impact of context effects on item parameter estimation, this silence is small comfort given the evidence that such effects do exist. In fact, IRT pre-equating under the new multistage rGRE takes the assumption of absence of context effects and resulting item change to its logical extreme in several ways. First, under IRT pre-equating every item is effectively an anchor, broadening the domain of concern. Second, immediate scoring provides no opportunity or mechanism for evaluating and confirming the absence of item change prior to reporting. Although evidence of score scale drift can and will be continuously gathered with the rGRE, this finding will make problems visible only in retrospect rather than in real-time as scores are being reported. Finally, whereas the assembly and equating of conventional test forms requires only that items remain stable across a handful of administration occasions and test-form contexts, the multistage rGRE will reuse items across numerous occasions and as part of a wide variety of test forms.

It is helpful at this point to describe the most likely model that the rGRE will adopt for pretesting and IRT calibration of new items, for linking these new calibrations to the operational scale, for assembling new multistage test forms, and for ensuring that scores on these forms are comparable with one another. The specific assumptions that are made regarding item change and the actions that are proposed for limiting change will be made clearer in the process.

Pretesting

The pre-equated nature of the rGRE means that all items must be pretested and calibrated before they can be used operationally. The question is whether IRT parameter estimates are tied to the specific context in which an item was pretested or generalized to remain valid across the range of contexts that an item may appear in once operational. Two approaches can be considered for producing item parameter estimates that are less context-bound. The first borrows a standard notion from experimental design, where potentially confounding factors that cannot be easily controlled are instead randomized. The idea is to pretest items in an essentially random context. That is, an unscored pretest item could appear anywhere in a test, surrounded by any combination of other items. Should randomization prove effective, it has the advantage of being fairly easy and convenient to implement in practice. The second, far less convenient approach would be to attempt to actually control context effects by pretesting and operationally administering items in intact and unchanging units or bundles.

Operational Item Use

Once calibrated, new items are subject to become productive members of operational forms. Under a multistage adaptive test, items will be bundled into the testlets that are successively selected and administered so as to comprise the most appropriate test form for each examinee. As far as context is concerned, several generalizations can be made. The first is that items will tend to be bundled with items of like difficulty. Testlets will therefore usually be more homogeneous in difficulty than are conventional test forms. Testlets will also most likely be administered intact or in their entirety to each examinee who receives them, forming a sort of context. However, it is possible to again resort to randomized presentation order—this time within testlets—to attempt to minimize the effects of specific context (Pomplun & Ritchie, 2004). Finally, the nature of the adaptive process means that very easy and very difficult items (or testlets) are somewhat more likely to appear near the end of an exam or timed section than

near the beginning. Because it takes a while for examinees to establish themselves as requiring extremely easy or difficult items, administration of such items generally occurs late in the exam. However both the Verbal and Quantitative measures will likely be divided into several separately timed sections, somewhat confusing the notion of what constitutes late in the exam.

Pre-Equating

As noted above, the rGRE will assemble and administer very large numbers of distinct test forms. In fact, any given administration occasion (a single day, or even part of a single day) will be supported by multiple unique (but overlapping) forms and the particular forms in use will change across occasions. The adaptive nature of administration means that different forms may differ dramatically in level of difficulty (although each will conform to content and other substantive specifications). It is this profusion of test forms, more than the requirement of immediate score reporting, that dictates IRT-based pre-equating. To be effective, IRT pre-equating requires that pretest-based item parameter estimates sufficiently characterize examinee performance once items become members of operational testlets. If scoring is by pre-equated number-correct rather than by IRT proficiency estimates, then pretest calibrations do not necessarily have to predict operational performance accurately for *each* item but only in the aggregate, averaged across items in a testlet or even an entire form. This requirement is significantly weaker.

All of the concerns identified above make it essential that the potential for context effects negatively impacting the rGRE be examined and evaluated. We review below some of the previous research on context effects and its implications for the design as well as the scoring and equating of the rGRE. Several new, more pointed studies intended to determine how context effects are best coped with in the specific context of the rGRE are then described.

3. Previous Research

Context effects have been extensively studied, with the following representing only the barest overview of existing research. However, two conclusions seem clear. First, compelling evidence shows that item performance can and does change from one fixed context to another (Dorans & Lawrence, 1990; Haladyna, 1992; Harris, 1991; Leary & Dorans, 1985). Although it seems likely that item position in the test is the dominant driver of change, several of the studies

above found performance differences even with position held relatively fixed while only the nature of surrounding items was varied.

A second clear conclusion is that IRT pre-equating, at least as evaluated, is not as effective as good, old-fashioned, post-administration equating by either IRT or conventional means (Eignor, 1985; Kolen & Harris, 1990). These studies concluded that item context (most notably item position) as well as the effects of multidimensionality might have contributed to the poorer performance of the item-pre-equating design. This result is unsurprising given that the pre-equating design makes stronger assumptions than does post-administration equating and does not generally allow these assumptions to be verified prior to score reporting.

These results present a stern warning that context effects cannot be safely ignored and that IRT pre-equating cannot be carelessly applied. However, the extent to which these difficulties impact the rGRE and the degree to which they can be controlled remains to be determined in greater detail. The results of a series of studies intended to do so are presented below.

4. Examining Context Effects in a GRE Environment

Item context effects can be quite situation specific, subject to the characteristics of the items, test forms, administration policies, and examinee population associated with a particular testing program. For example, a test administered under shorter time limits is more likely to show position effects than that same test administered under more generous timing. Item types and formats can also interact with time limits and create position effects. For example, the GRE Verbal measure includes "long" passage sets that require examinees to read a lengthy passage before answering the attached comprehension items. Positioning such a set near the end of a timed section may well exacerbate position effects as compared to concluding the test with a series of shorter, discrete items. Examinee populations (and sub-populations) are also likely to have a clear impact. The same test administered under the same conditions may be much more speeded for some examinee groups than for others.

Context effects are therefore ideally evaluated under circumstances that closely approximate the testing program in question. As detailed below, this evaluation was possible only to an extent with the studies conducted to examine the impact of context effects on the scoring and equating of the multistage GRE. Although data collection constraints allowed some of the conditions that will prevail under the rGRE to be replicated, others could not be.

Context has been defined above as encompassing both item position and the nature of surrounding items. Given the design of the rGRE, item position effects are much more easily controlled in calibration than are the effects of surrounding items. For example, consider the strategy of pretesting items in random locations throughout the test. Based on pretest estimates of item performance, one can assign an item to an appropriate testlet. However, one would also know where that testlet was destined to appear (early or late) as part of an operational form. Pretest data could then be accordingly resampled, with calibrations based only on pretest responses that occurred somewhere near where the item and its testlet will appear operationally. Although this would be inconvenient, it is at least tractable.

Controlling the exact nature of surrounding items is a good deal less convenient. Here, it would be necessary to form items into testlets based on (perhaps preliminary) pretest data and then submit the intact testlet to a second round of pretesting as a unit. Only data so collected could then be used for calibration. It therefore seems important to know whether position or the nature of surrounding items is the primary driver of context effects. Since prior research indicates strongly that position is dominant, two research questions formed the central focus of the current investigation:

Question 1: Are item position effects evident in linear test forms administered to GRE examinees?

Question 2: Are item position effects likely to pose a particular challenge to multistage tests?

As will be seen, restrictions on data collection did not permit either of these questions to be answered to complete satisfaction.

5. Data Collection Events

Data were collected across three occasions to inform answers or partial answers to the research questions posed. All three data collections employed the variable section of the computer adaptive test (CAT) version of the GRE. The variable section is embedded within each examinee's operational test and is typically used for pretesting new items. Each examinee receives a single variable section, which may contain either verbal or quantitative items. Because the variable section is not specifically identified to examinees as an unscored component of their exam, high motivation is ensured. However, the drawback is that the variable section must look

to examinees like an operational section. This precludes administering the new GRE item types, allowing the use of calculators, manipulating time limits or test length, or permitting the review and/or revision of previous responses. Because the rGRE is expected to introduce new item types, allow use of calculators, have different test lengths and time limits, and permit examinees to freely navigate within timed test sections, these limitations are not inconsiderable. Great caution is therefore required in extending inferences from this study to the rGRE.

Data Collection 1

Items and test forms. Sets of 28 quantitative items and 30 verbal items were arranged into linear (nonadaptive) test forms. These items were drawn from the current CAT GRE items banks and were judged as broadly representative of those banks. Each item set was required to meet the content specifications currently applied to operational CAT GRE exams. As will be seen later, these items were subsets of larger collections that were administered to approximate a multistage test. That study will be described below as data Collection 3.

Conditions. Linear test forms were administered to 5,693 GRE examinees as part of the variable section attached to GRE exams delivered in July and August of 2008. As noted, the variable section was administered under the same conditions as an operational section of the same content. The verbal variable section was therefore timed at 30 minutes; examinees were given 45 minutes to complete a quantitative section. Because the variable section was administered as a component of an actual GRE exam, each examinee's operational GRE scores were also available.

Design. Verbal and quantitative item collections (comprising 30 and 28 items, respectively) were administered in each of several scrambled orders. The quantitative items were administered in 13 distinct orderings; the verbal items were presented in seven different orders. The scrambled orders were designed such that each item appeared with equal frequency in each of several general locations throughout the test. The different item orderings are shown in Tables 1 and 2.

Because only a single variable section was administered to each examinee, the total sample was randomly split in two, with half being administered a verbal variable section while the second half took a quantitative section. Within each of these halves, the sample was randomly subdivided still further and apportioned across the different item orderings that were available. One of the orderings within each content area was designated as a base and

administered to a much larger sample than were the others. Sample sizes by ordering are presented in Table 3.

Table 1

Quantitative Section Item Positions by Ordering

Base	2	3	4	5	6	7	8	9	10	11	12	13
1	27	25	23	21	19	16	14	12	10	7	5	4
2	26	22	18	14	9	4	28	24	20	15	11	7
3	1	27	25	23	21	18	16	14	12	9	8	6
4	28	24	20	16	12	6	2	26	22	17	13	9
5	3	1	27	25	23	20	18	16	14	12	10	8
6	2	26	22	18	14	9	4	28	24	19	15	11
7	5	3	1	27	25	22	20	18	17	14	12	10
8	4	28	24	20	16	11	7	2	26	21	17	13
9	7	5	3	1	27	24	22	21	19	16	14	12
10	6	2	26	22	18	13	9	5	28	23	19	15
11	9	7	5	3	1	26	25	23	21	18	16	14
12	8	4	28	24	20	15	11	7	3	25	21	17
13	11	9	7	5	3	1	27	25	23	20	18	16
14	10	6	2	26	22	17	13	9	5	28	23	19
15	13	11	9	7	6	3	1	27	25	22	20	18
16	12	8	4	28	24	19	15	11	7	2	26	21
17	15	13	11	10	8	5	3	1	27	24	22	20
18	14	10	6	2	26	21	17	13	9	4	28	24
19	17	15	14	12	10	7	5	3	1	26	24	22
20	18	16	15	13	11	8	6	4	2	27	25	23
21	16	12	8	4	28	23	19	15	11	6	2	26
22	20	19	17	15	13	10	8	6	4	1	27	25
23	19	14	10	6	2	25	21	17	13	8	4	28
24	23	21	19	17	15	12	10	8	6	3	1	27
25	21	17	12	8	4	27	23	19	15	10	6	2
26	22	18	13	9	5	28	24	20	16	11	7	3
27	25	23	21	19	17	14	12	10	8	5	3	1
28	24	20	16	11	7	2	26	22	18	13	9	5

Note. Interpretation: Item 1 in the base ordering was moved to position 27 in ordering 2, position 25 in ordering 3, and so on.

Table 2

Verbal Section Item Positions by Ordering

Base	2	3	4	5	6	7
1	30	26	21	17	10	6
2	24	15	9	15	22	12
3	21	7	28	29	9	17
4	18	30	16	21	14	25
5	1	22	25	2	26	10
6	2	23	26	3	27	11
7	25	11	14	19	4	16
8	22	4	20	26	13	24
9	6	1	30	10	21	29
10	3	27	18	24	8	15
11	26	16	3	5	17	20
12	27	17	4	6	18	21
13	28	18	5	7	19	22
14	29	19	6	8	20	23
15	10	8	24	30	25	28
16	7	5	1	14	3	19
17	4	24	22	28	12	27
18	14	12	10	12	24	5
19	11	9	29	1	30	26
20	8	2	8	18	7	4
21	5	28	27	16	16	9
22	19	20	15	9	29	30
23	15	13	12	22	1	2
24	16	14	13	23	2	3
25	12	6	19	4	11	8
26	9	25	7	20	23	14
27	23	21	17	13	6	1
28	20	10	2	27	15	7
29	17	3	23	11	28	13
30	13	29	11	25	5	18

Note. Interpretation: Item 1 in the base ordering was moved to position 30 in ordering 2, position 26 in ordering 3, and so on.

Table 3

Data Collection 1, Examinee Count by Ordering

Verbal measure		Quantitative measure	
Ordering	Count	Ordering	Count
1 (Base)	1,315	1 (Base)	1,315
2	200	2	114
3	198	3	117
4	219	4	115
5	195	5	114
6	212	6	121
7	215	7	100
		8	120
		9	116
		10	127
		11	86
		12	95
		13	98

Data Collection 2

Items and test forms. An additional 80 quantitative items were divided into 3 sets of 28 items each; with 4 of the 80 items appearing in 2 different sets. Each of the 3 sets was then arranged into 13 distinct orders, following the same pattern described for the first data collection event.

Eighty-four verbal items were similarly distributed across 3 sets, each of which was ordered in 7 ways. Among the 84 items, 6 were administered in 2 different sets. As with the first data collection, both the quantitative and verbal items sets were selected to meet current content specifications so that all appeared indistinguishable from operational sections.

Conditions. The quantitative and verbal item sets and orderings were administered to 11,245 examinees tested during September and October of 2008. All of the conditions for these administrations parallel those described for the first data collection.

Design. The various item sets and orderings (39 for quantitative and 21 for verbal) were again randomly spiraled across examinees, with each item appearing in various locations throughout the test section with roughly equal frequency. Examinee counts by item set and ordering are shown in Table 4.

Table 4

Data Collection 2, Examinee Count by Form/Ordering

Verbal measure		Quantitative measure	
Set/ordering	Count	Set/ordering	Count
1 / 1	274	1 / 1	145
1 / 2	273	1 / 2	140
1 / 3	262	1 / 3	123
1 / 4	237	1 / 4	133
1 / 5	261	1 / 5	142
1 / 6	241	1 / 6	122
1 / 7	258	1 / 7	135
2 / 1	273	1 / 8	129
2 / 2	246	1 / 9	150
2 / 3	279	1 / 10	132
2 / 4	295	1 / 11	153
2 / 5	272	1 / 12	136
2 / 6	267	1 / 13	145
2 / 7	257	2 / 1	163
3 / 1	261	2 / 2	127
3 / 2	288	2 / 3	145
3 / 3	286	2 / 4	140
3 / 4	259	2 / 5	137
3 / 5	267	2 / 6	151
3 / 6	266	2 / 7	139
3 / 7	285	2 / 8	121
		2 / 9	146
		2 / 10	161
		2 / 11	159
		2 / 12	129
		2 / 13	129
		3 / 1	142
		3 / 2	165
		3 / 3	138
		3 / 4	154
		3 / 5	130
		3 / 6	143
		3 / 7	136
		3 / 8	137
		3 / 9	143
		3 / 10	138
		3 / 11	153
		3 / 12	119
		3 / 13	136

Data Collection 3

Items and test forms. Although the current test delivery mechanism does not directly support multistage tests, it can be manipulated into administering a reasonable approximation of one. This approximation was done by combining all of the items administered in the first two data collections as the components of a multistage test. Figures 1 and 2 show how the various stages and modules of the multistage test were organized and labeled.

Each examinee began with Module 103, which was comprised of 8 items of moderate difficulty for both measures. Contingent on their performance on Stage 1, examinees were routed to any of five available second-stage modules. These were labeled 201 through 205 and ranged from quite easy to quite difficult. Again, contingent on performance, examinees were similarly routed through Stages 3 and 4, both of which also contained five modules of varying levels of difficulty. Although this multistage test design does not directly parallel what has been adopted for the rGRE, its impact on examinees should be similar.

Conditions. Verbal and quantitative multistage tests were administered in the variable section to 2,947 and 2,916 examinees, respectively, who tested in March and April of 2008. Note that this testing actually preceded administration of the scrambled item sets from which the multistage tests were built. However, this fact is unimportant to either the outcome or the description of our investigation of context effects. Since the variable section was employed for administration, the multistage tests again had to emulate operational sections in terms of test length, timing, and item composition.

Design. Either the verbal or the quantitative multistage test was randomly selected for each examinee. Each multistage test resulted in the delivery of 4 of the 16 available modules, with different examinees taking different combinations, depending on their performance. Table 5 shows the numbers of examinees administered each module. Because these numbers are aggregated across routes, the table does not convey the full complexity of the data collection design.

6. Analyses and Results

Analysis of data from the three collection events was organized around the two main research questions: are item position effects evident in GRE data and are they likely to pose a particular challenge to a multistage test? The analyses informing answers to these questions will be described in turn.

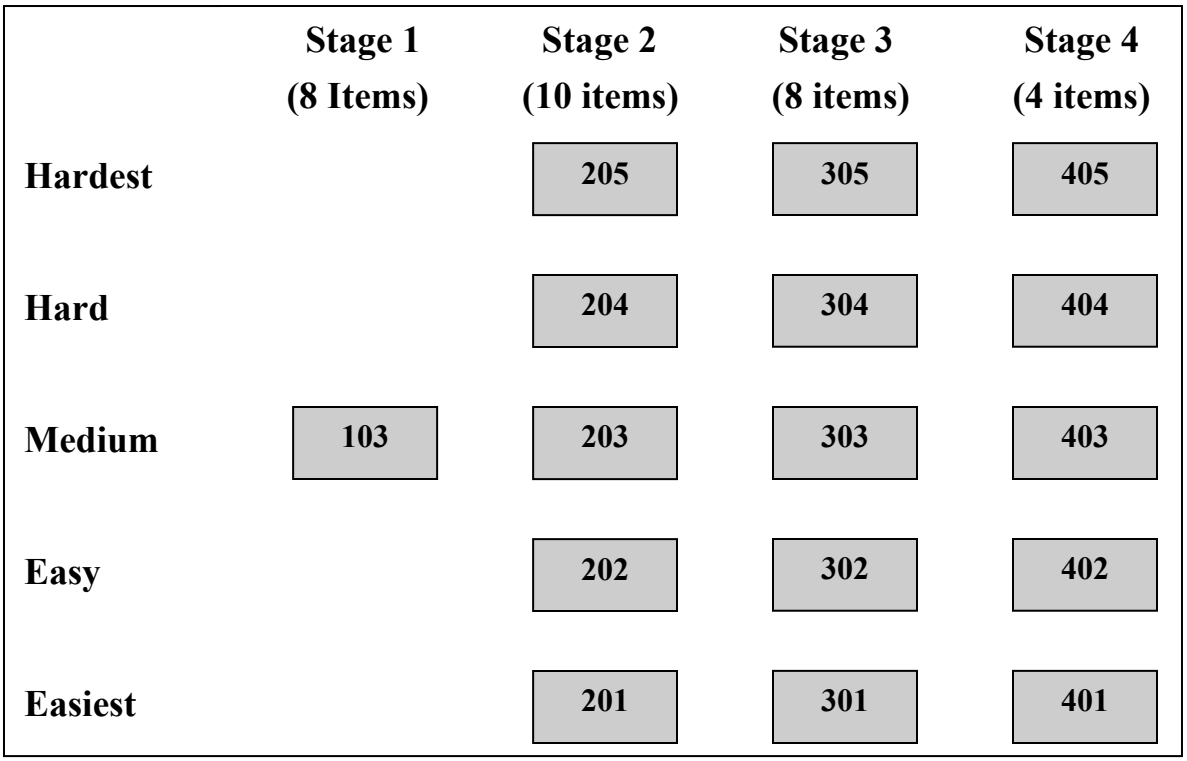

Figure 1. **Multistage test structure for verbal measure.**

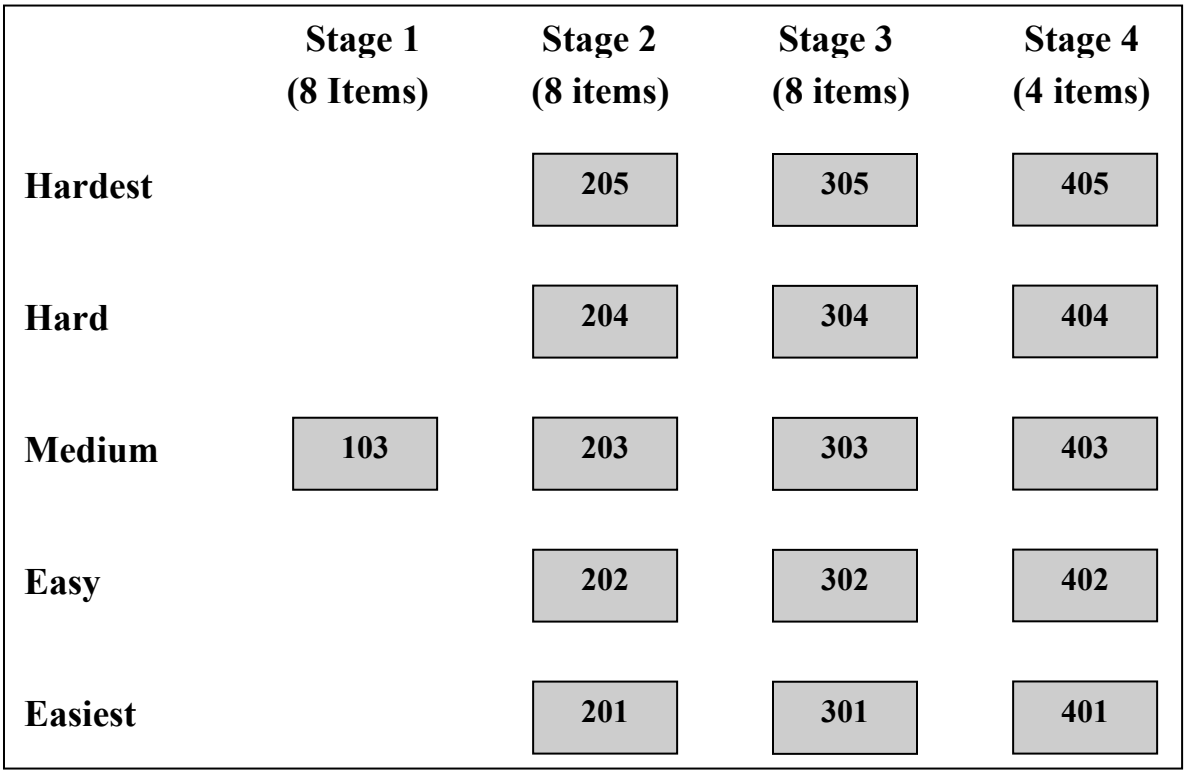

Figure 2. **Multistage test structure for quantitative measure.**

Table 5

Data Collection 3, Examinee Count by Test Module

Verbal measure		Quantitative measure	
Module	Count	Module	Count
103	2,623	103	2,432
201	675	201	468
202	723	202	416
203	578	203	539
204	381	204	412
205	266	205	597
301	631	301	363
302	720	302	475
303	748	303	660
304	413	304	579
305	111	305	355
401	608	401	371
402	751	402	517
403	796	403	712
404	389	404	551
405	79	405	281

Question 1: Are Item Position Effects Evident in Linear GRE Data?

The scrambled test forms administered during the first data collection were specifically designed to address this question. Recall that linear (nonadaptive) tests were administered in any of 13 distinct orderings for the quantitative measure and 7 orderings for verbal. Among each set of orderings was a base ordering for which sample sizes were much larger.

Data screening. Prior to formal analyses, each dataset was subjected to screening to identify and eliminate examinees who either failed to achieve some minimal level of performance or appeared to be responding in an unusual fashion. Examinees who performed at less than chance level on the variable section or whose variable section performance was much different than their operational GRE score would suggest were therefore removed.[1] To simplify analyses, examinees who failed to complete the variable section were also eliminated. Less than 10% of the data records were eliminated as a result of all of this screening. Qualifying sample sizes for each item ordering are shown in Table 3.

Exploratory analyses. Many of the analyses that follow depend strongly on random equivalence of the examinee groups administered each item ordering. Although this analysis was theoretically ensured by randomly spiraling orderings across reasonably large samples, the success of spiraling was nonetheless evaluated. Operational GRE scores proved important in this regard, as they served as an independent measure of the proficiency level of examinee groups. For both verbal and quantitative measures, neither analysis of variance (ANOVA) nor pair-wise t-tests revealed any significant differences in operational score distributions across examinee groups. This finding suggests that any observed differences in performance across item orderings is in fact the result of the ordering rather than due to underlying group differences.

Item difficulty comparisons across orderings. The first and simplest analysis of position effects was to compare whether an item's difficulty, here measured by *p values* or proportions of correct responses, differed depending on where the item appeared in the test section. Note that item *p* values are proportions, so the variance of item *p* values varies across different true item *p* values. Because statistical analyses often rely on the assumption of homogeneous variances, *p* values were transformed to a scale on which variances are more stable and comparable (Weisberg, 1985).

The standard variance-stabilizing transformation for proportions is the arcsine transformation, where $t(p) = \sin^{-1}\left(\sqrt{p}\right)$, with *p* the *p* value for a given item in a given ordering. This transformation was applied to each item *p* value in each ordering.

Introducing some notation, let p_{Bi} be the *p* value for item *i* in the base, or large sample ordering. Similarly, let p_{Si} be the *p* value observed for item *i* in each of the scrambled orderings, where $2 \le s \le 7$ and $2 \le s \le 13$ for verbal and quantitative, respectively. Then the difference $d_s p_i = t\left(p_{si}\right) - t\left(p_{Bi}\right)$ is the change in (transformed) *p* value observed after moving item *i* from its position in the base ordering to its position in ordering *s*.

Further, let x_{Bi} and x_{si} be the positions item *i* takes in either the base or additional orderings. Then the corresponding difference between these positions can be denoted $d_s x_i = x_{si} - x_{Bi}$. This difference is the distance that item *i* was moved from its position in the base ordering to its position in ordering *s*. Note that this distance is positive or negative,

depending on whether the item was moved backward or forward from its position in the base ordering.

Since analyses of *p* value change due to shift in position were certain to be idiosyncratic and unstable at the individual item level, results were aggregated across items that shifted similar distances. Table 6 describes exactly how this aggregation was done for each measure. For example, for the quantitative measure, 42 *p* value differences that involved shifts backward from the base to the scrambled ordering of distances amounting to more than half the test length were grouped together. Similarly, 62 differences of between 7 and 14 positions backward were grouped together. The resulting aggregation groups are termed *gdx* in the table and in the discussion that follows.

Each aggregation group, or *gdx,* essentially defines a distribution of differences, comprised as it is of differences associated with various items that shifted similar distances in the administration order. Although traditional hypothesis tests were applied to discover whether distributions differed significantly across shifts of various distances, the results of this comparison are perhaps best presented graphically, as displayed in Figures 3 and 4.

Table 6

Grouping of Item Movement Distances

	Quantitative measure		Verbal measure		
gdx	Range of *dx*	Count	Range of *dx*	Count	Description
1	[-26,-15]	42	[-26,-16]	24	Moved forward more than half the test length.
2	[-14 , -7]	62	[-15 , -7]	37	Moved forward between a quarter and half of the test length.
3	[-6 , -1]	72	[-6 , -1]	36	Moved forward less than a quarter of the test length.
4	0	28	0	30	Same position as in the base ordering.
5	[1 , 6]	51	[1 , 6]	21	Moved backward less than a quarter of the test length.
6	[7 , 14]	67	[7 , 15]	30	Moved backward more than a quarter but less than half of the test length.
7	[15 , 26]	42	[16 , 29]	32	Moved backward more than half of the test length.

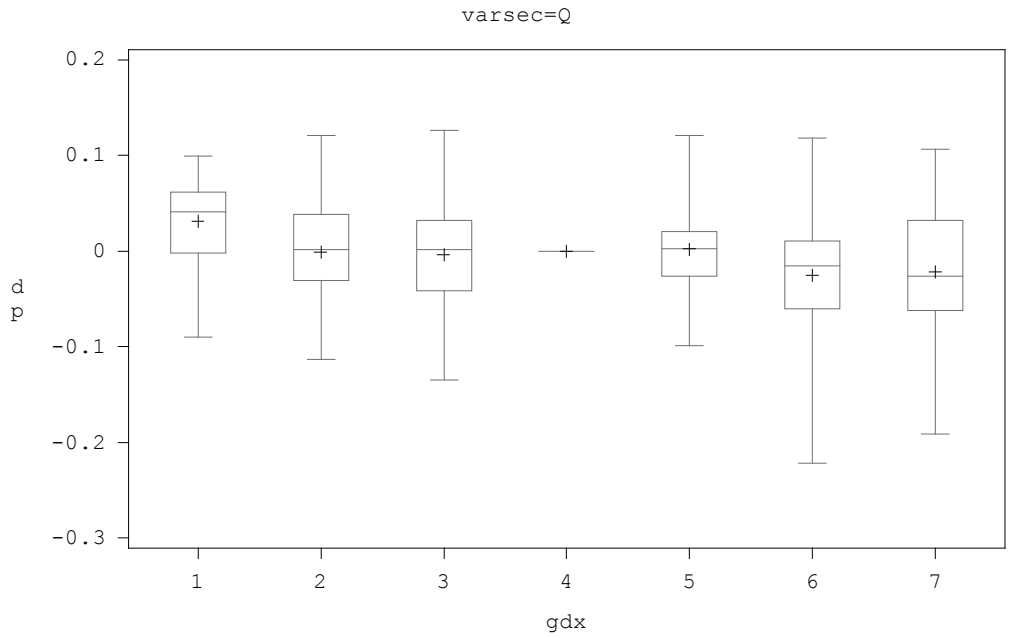

Figure 3. **Box plots for *p* value differences due to item position shift: quantitative measure. Mean differences are significantly nonzero for *gdx* = 1, 6, and 7.**

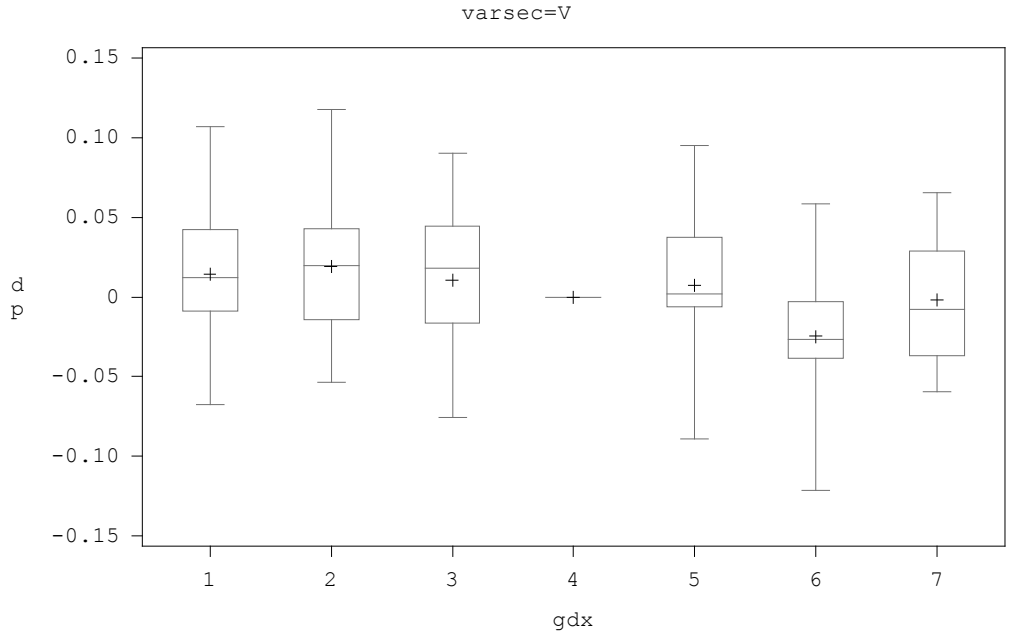

Figure 4. **Box plots for *p* value differences due to item position shift: verbal measure. Mean differences are significantly non-zero for *gdx* = 2 and 6.**

Figures 3 and 4 describe each distribution by a box plot (Tukey, 1977) that characterizes various distribution features as follows:

- The central box encloses the interquartile range, demarked by the 25th and 75th percentiles of the distribution.
- The solid line bisecting each box indicates the median, or 50th percentile.
- The cross (+) locates the mean.
- The "whiskers" attached to each side of the box extend to the observed minimum and maximum values.

The center box anchors this display by representing items that neither shifted in position nor changed in performance. Boxes left of center characterize items that moved forward while boxes right of center depict items that moved backward. Similarly, *p* value differences on the positive side of the vertical center indicate that items became slightly easier when shifted while differences on the negative side of the vertical center indicate items became more difficult. A bit of imagination allows a trend to be detected from both verbal and quantitative, with items becoming easier when moved forward and more difficult when moved backward, relative to the difficulty and position in the base ordering. This finding is consistent with the vast majority of prior research on position effects. More formal hypothesis testing confirmed that the most extreme moves forward or backward (e.g., $gdx = 1$ or $gdx = 7$) tended to produce significant *p* value differences.

Conditioning on item difficulty. A weakness of the above analysis is that items of varying (and perhaps very different) difficulty levels were grouped together in each of the shift-distance strata. This grouping is problematic because we expect the size of a *p* value change to be at least in part related to item difficulty itself (the effects of the variance-stabilizing transformation notwithstanding). Part of the problem is related to the difficulty measure being effectively bounded at both extremes. Difficult items can become only so much more difficult if moved back in the section while easy items can become only so much easier if moved forward.

Items were therefore sorted into three groups based on their difficulty (as computed across all orderings). Logically enough, these three item groups were labeled as easy, medium, and hard. The box plots described above were produced again, this time for each of the three item difficulty groups. These plots are shown in Figures 5 to 10. It appears that the relationship between shift distance and difficulty change is stronger for quantitative than for verbal, and it is especially pronounced when the easiest items are moved back in the test section.

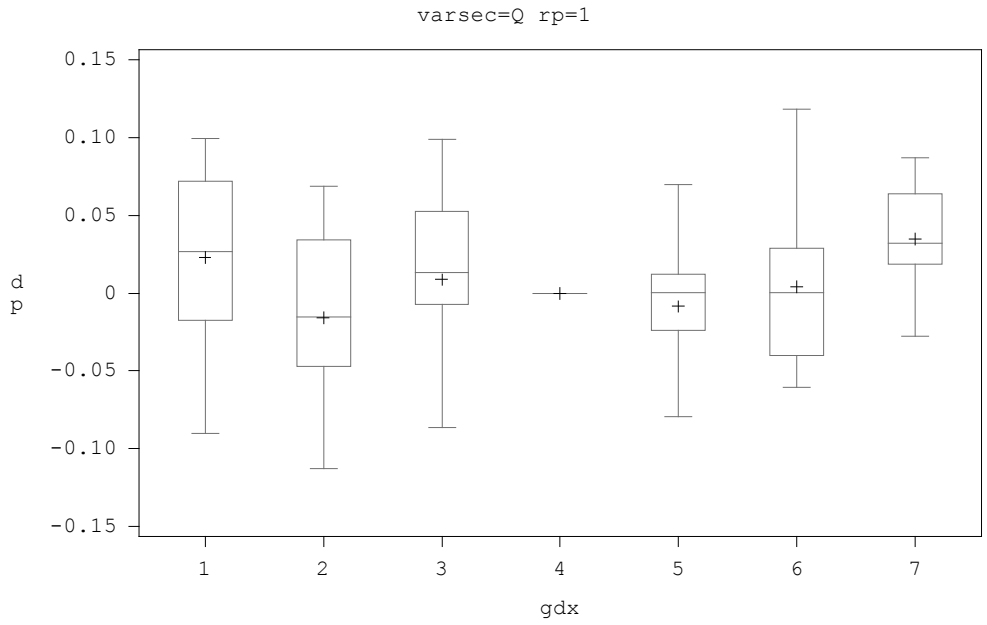

Figure 5. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: quantitative, hard**

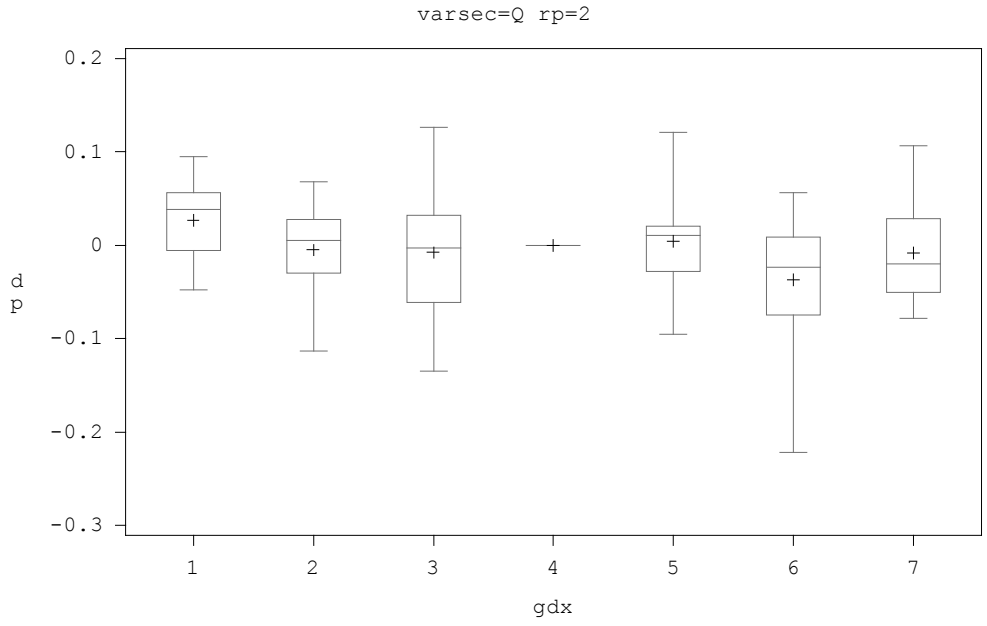

Figure 6. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: quantitative, medium. The *dp* is significantly non-zero for *gdx* = 1 and 6**

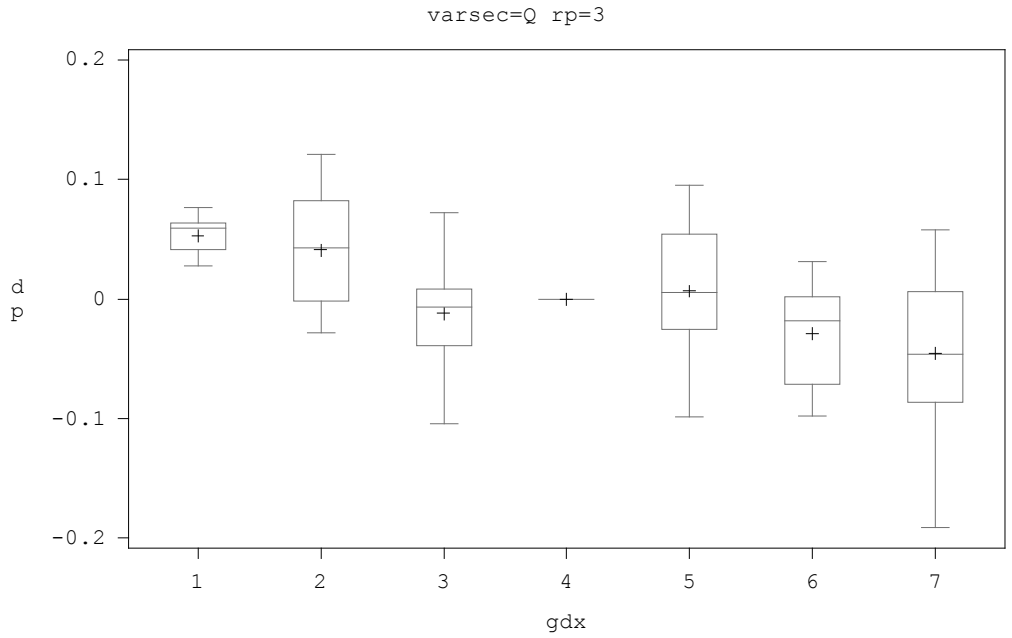

Figure 7. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: quantitative, easy. The *dp* is significantly non-zero for *gdx* = 1, 2, 6, and 7**

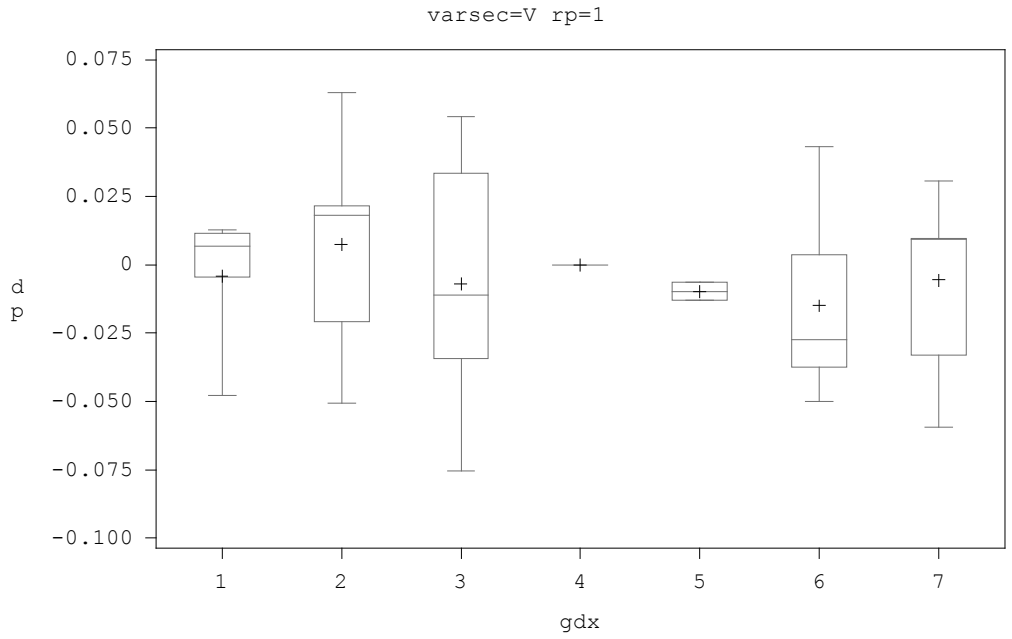

Figure 8. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: verbal, hard**

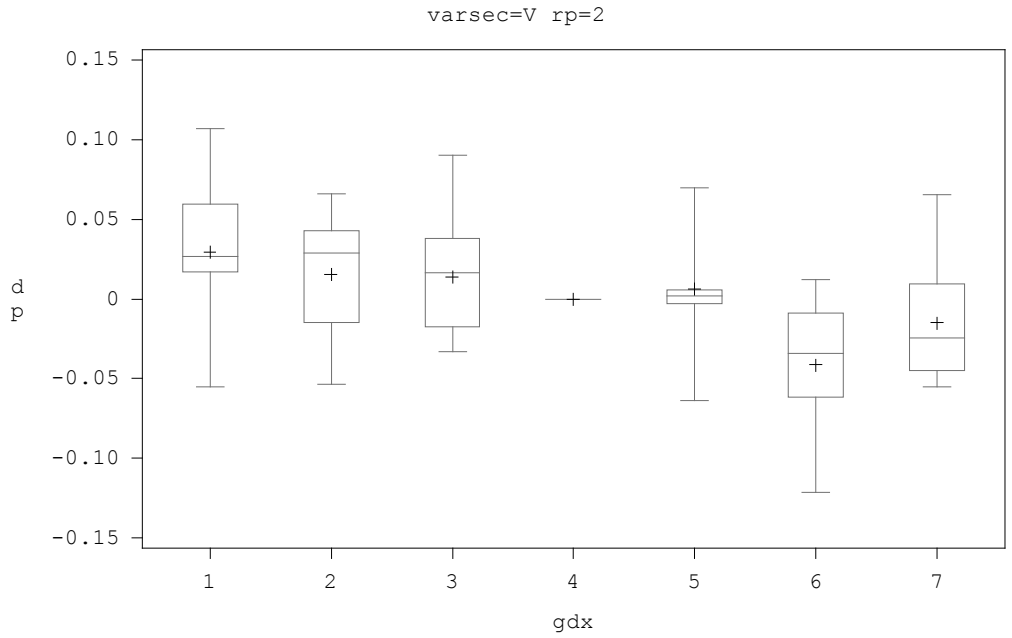

Figure 9. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: verbal, medium. The *dp* is significantly non-zero for *gdx* = 1 and 6**

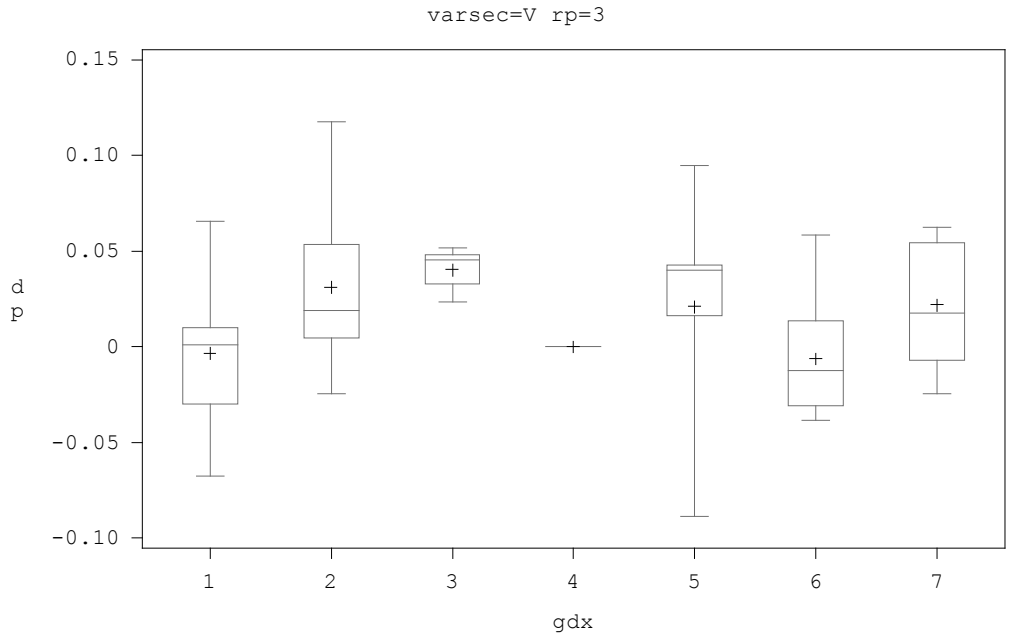

Figure 10. **Box plots for *p* value differences due to item position shift, grouped by item difficulty: verbal, easy. The *dp* is significantly non-zero for *gdx* = 2 and 3**

Conditioning on item type. Item type or format is another potential factor in the extent of difficulty change following a shift in position. Most relevant to GRE is the distinction between *discrete* or individual, self-contained items and passage-based *item sets*. The latter appear on the verbal measure and require examinees to read a passage of several hundred words before answering two or four associated questions. Passage-based sets can be particularly vulnerable to test speededness if they appear late in a section when examinees are likely to feel time pressure.

Figures 11 and 12 show the by now familiar box plot by item shift array, this time for the Verbal measure only but separately for discrete and passage-based items. It is evident that the relationship between shift distance and difficulty change is stronger for passage-based items than for discrete items.

Conditioning on consequential movement. A consequential movement refers to a shift from or to the last three slots of a test. Because speededness was likely to affect the extent of difficulty change following a shift in position, items associated with consequential movements were isolated and further analyzed. Generally, the results reveal the same pattern as those in Figures 3 to 12. More striking relationship between shift distance and difficulty change is found for quantitative items with consequential movements. Thus test speededness is likely to affect the difficulty of quantitative items more than the difficulty of verbal items.

Logistic Modeling

The same results presented above in an exploratory and graphical fashion were replicated through a more formal modeling approach. With binary data such as correct or incorrect responses, the model of choice is *logistic regression* and was applied as follows. Let $\pi(x)$ be the expected proportion of correct responses, or the expected *p* value, for a certain item at position *x* in the test section. Consider the model

$$\text{logit}\left(\pi(x)\right) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x,$$

where α is the intercept, and β is the rate of change of the observed item difficulty as the position of the item changes in the test section.

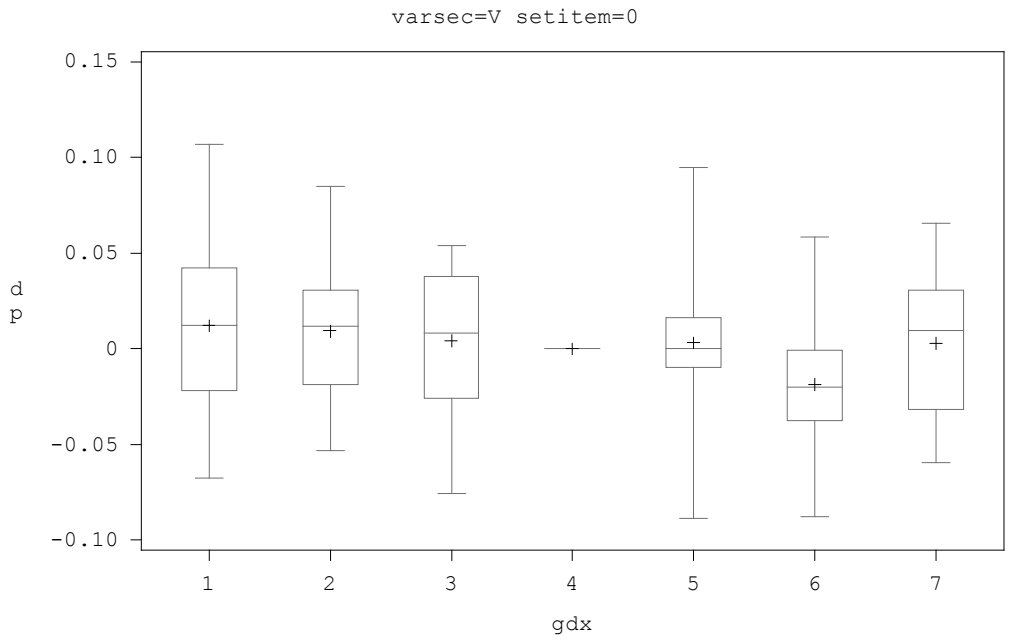

Figure 11. **Box plots for *p* value differences due to item position shift, grouped by item type: verbal, discrete items.**

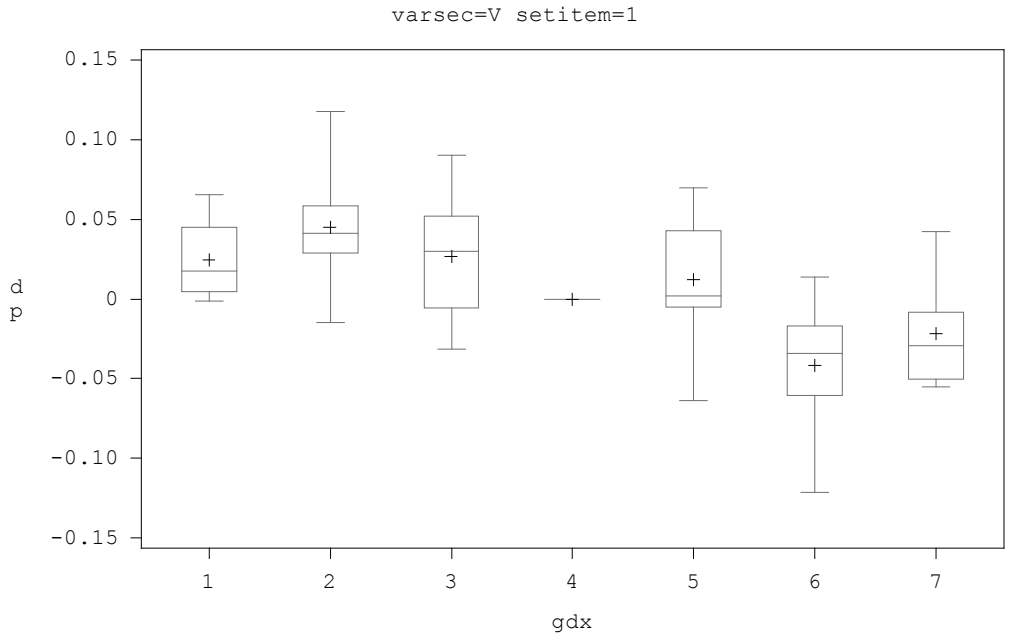

Figure 12. **Box plots for *p* value differences due to item position shift, grouped by item type: verbal, passage-based items. The *dp* is significantly non-zero for *gdx* = 2 and 6.**

Under this model, the (null) hypothesis that item difficulty is unaffected by shift in position can be evaluated by formal significance tests. The likelihood-ratio (LR) test of the null hypothesis $\beta = 0$ is particularly to the point (Agresti, 1996). Results of logistic regression modeling and significance testing are summarized in Tables 7 and 8. For the quantitative measure, 8 of the 28 items showed significant position effects. Only 4 of the 30 verbal items showed significant effects at the same level. Characteristics of the identified items were reviewed and found unremarkable. Most were discrete (as are the preponderance of items) and were of medium or easier difficulty levels.

Item Response Theory (IRT) Modeling

Item response theory can offer a more refined look at position effects (Lord, 1980). The two-parameter logistic (2PL) model was first fit to data from the base orderings of the verbal and quantitative measures. The 2PL model was used in all of our analyses, both item calibration and proficiency estimation, as it will be the future practice for the rGRE. Position effects can then be measured by how well these parameter estimates predicted the responses observed to the same items in the various scrambled orderings.

The quality of predictions can be assessed through any of a variety of residual measures. The particular measure applied here was computed as follows. First, proficiency estimates were produced for all examinees, based on the item parameters estimated from the base orderings. Proficiencies were then grouped into *m* strata. The *j*th stratum, $1 \le j \le m$, contained N_j examinees. Consider now a particular item *i* administered in a particular position in a given scrambled ordering. Let P_{ij} denote the observed proportion of correct responses on this item among the examinees in proficiency stratum *j*, $1 \le j \le m$. Further, let $E\left(P_{ij}\right)$ be the expected proportion of correct responses on this item for the examinees in this stratum. This proportion is given by:

$$E\left(P_{ij}\right) = \frac{1}{N_j} \sum_{k=1}^{N_j} \frac{1}{\left(1 + exp\left(-1.7 a_i \left(\hat{\theta}_k - b_i\right)\right)\right)},$$

where $\hat{\theta}_k$ is the proficiency estimate for examinee *k*, and a_i and b_i are the parameter estimates for the item in question (again, from the base ordering).

Table 7

Estimates of β and Significance: Quantitative

Item	β	SE	LR statistic	LR *p* value
Q-1	-0.002	0.005	0.100	0.753
Q-2	-0.006	0.005	1.610	0.205
Q-3	-0.001	0.005	0.060	0.800
Q-4	-0.009	0.006	2.790	0.095
Q-5	0.001	0.007	0.030	0.855
Q-6	-0.032	0.007	19.310	<.0001*
Q-7	0.001	0.007	0.020	0.894
Q-8	-0.004	0.006	0.370	0.544
Q-9	0.012	0.006	3.700	0.055
Q-10	-0.013	0.007	4.180	0.041*
Q-11	-0.012	0.007	2.660	0.103
Q-12	-0.029	0.008	12.930	0.001*
Q-13	-0.001	0.007	0.000	0.944
Q-14	-0.017	0.008	5.290	0.021*
Q-15	0.001	0.007	0.020	0.874
Q-16	-0.001	0.007	0.030	0.858
Q-17	0.009	0.006	1.980	0.160
Q-18	-0.016	0.007	5.250	0.022*
Q-19	-0.001	0.006	0.010	0.941
Q-20	-0.003	0.006	0.160	0.690
Q-21	-0.019	0.007	7.460	0.006*
Q-22	0.001	0.006	0.050	0.832
Q-23	0.003	0.006	0.230	0.628
Q-24	-0.004	0.005	0.540	0.462
Q-25	-0.004	0.005	0.540	0.463
Q-26	-0.010	0.006	3.230	0.073
Q-27	-0.011	0.005	6.080	0.014*
Q-28	-0.011	0.005	4.590	0.032*

* p value < 0.05.

Table 8

Estimates of β and Significance: Verbal

Item	Estimate	SE	LR statistic	LR *p* value
V-1	0.000	0.004	0.000	0.945
V-2	0.006	0.007	0.820	0.365
V-3	-0.006	0.004	1.820	0.178
V-4	-0.006	0.004	1.890	0.169
V-5	0.001	0.005	0.020	0.891
V-6	-0.009	0.005	3.600	0.058
V-7	-0.012	0.007	3.090	0.079
V-8	-0.006	0.006	0.810	0.367
V-9	0.006	0.006	0.930	0.335
V-10	0.009	0.007	1.490	0.222
V-11	-0.006	0.008	0.540	0.462
V-12	-0.013	0.008	2.790	0.095
V-13	-0.025	0.007	12.780	0.001*
V-14	-0.010	0.007	2.060	0.151
V-15	-0.008	0.006	1.570	0.211
V-16	-0.008	0.008	1.030	0.311
V-17	-0.002	0.007	0.060	0.805
V-18	-0.003	0.009	0.080	0.779
V-19	-0.016	0.006	6.820	0.009*
V-20	-0.001	0.006	0.020	0.898
V-21	0.001	0.007	0.000	0.946
V-22	0.007	0.008	0.730	0.394
V-23	-0.004	0.006	0.630	0.428
V-24	-0.009	0.005	3.530	0.060
V-25	0.006	0.006	1.160	0.282
V-26	-0.004	0.006	0.390	0.530
V-27	-0.010	0.005	4.450	0.035*
V-28	0.000	0.006	0.000	0.953
V-29	-0.015	0.005	9.110	0.003*
V-30	0.005	0.005	1.180	0.278

***** *p* value < 0.05.

For item *i* in a given scrambled ordering, the residual r_i is then given as

$$r_i = \frac{\sum_{j=1}^{m} P_{ij} - \sum_{j=1}^{m} E\left(P_{ij}\right)}{\sqrt{\sum_{j=1}^{m} E\left(P_{ij}\right)\left(1 - E\left(P_{ij}\right)\right) / N_j}},$$

for $1 \le i \le 30(28)$ for verbal (quantitative). By the central limit theorem, this residual asymptotically follows a standard normal distribution. Concerns regarding choice of *m* are discussed by Hambleton, Swaminathan, and Rogers (1991). In our study, *m* = 10 was chosen for sample sizes greater than 300 (i.e., base orderings) and *m* = 5 was used for sample sizes less than 300 (i.e., scrambled orderings). These numbers were deemed appropriate based on the evaluation of the empirical distribution for each *i*.

If position has an effect on item performance, the residuals for an item should be positive when it is administered earlier in a scrambled order than in the base ordering. This residual would indicate that the item appeared easier in a scrambled ordering than the base ordering items parameter estimates predicted. Similarly, one would expect negative residuals for items that appeared later in a scrambled ordering than in the base.

Results are summarized by the box plots in Figures 13 and 14, which group the residuals for items in specific positions by the degree of shift from the base ordering. The same classification rules described in Table 5 were used to group residuals by the distance that items shifted from the base ordering. Each box in Figures 13 and 14 then characterizes the distribution of residuals for items shifted a similar distance. A clear trend is visible in both plots, with items shifted to earlier positions becoming easier while items shifted to later positions become more difficult. The clarity of these trends in comparison to those shown in Figures 3 and 4 implies that IRT-based analyses are indeed a more sensitive measure of position effects.

Mitigating Position Effects

The conclusion that position effects are evident in linear GRE forms begs the question of whether particular pretest strategies can minimize their impact on pre-equating. One such strategy is to pretest items in a variety of positions throughout a section rather than in a single, fixed position (as evaluated above). Happily, the design employed in the first two data collection

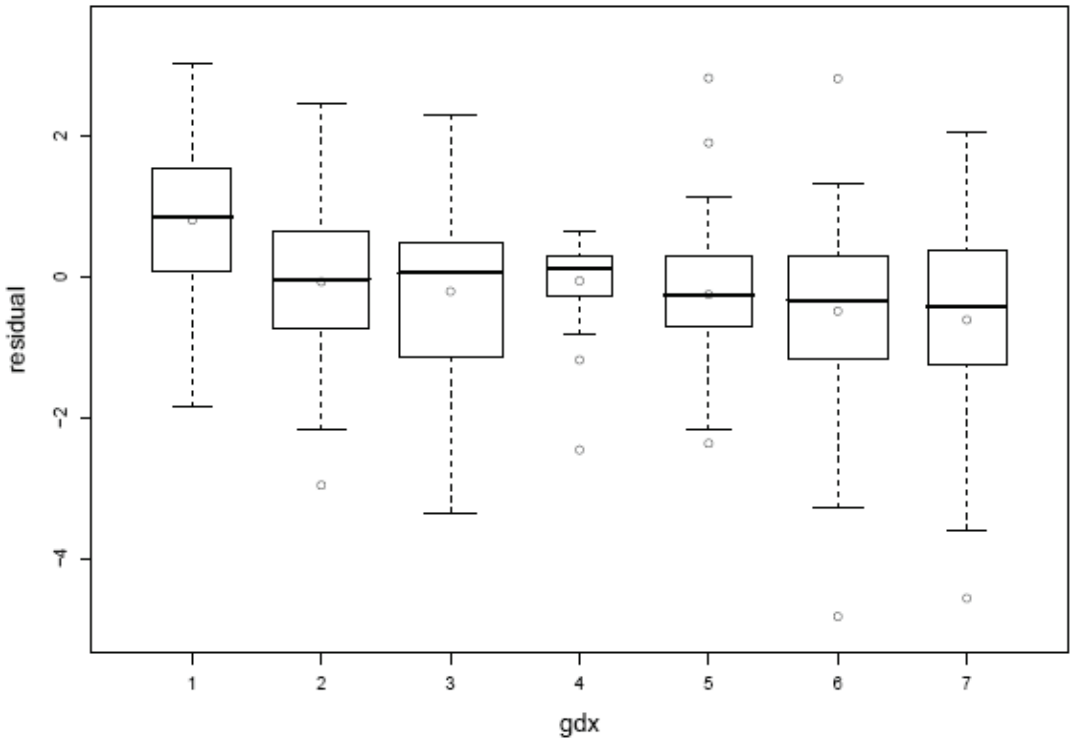

Figure 13. **Box plots for IRT residuals due to item position shift: quantitative.**

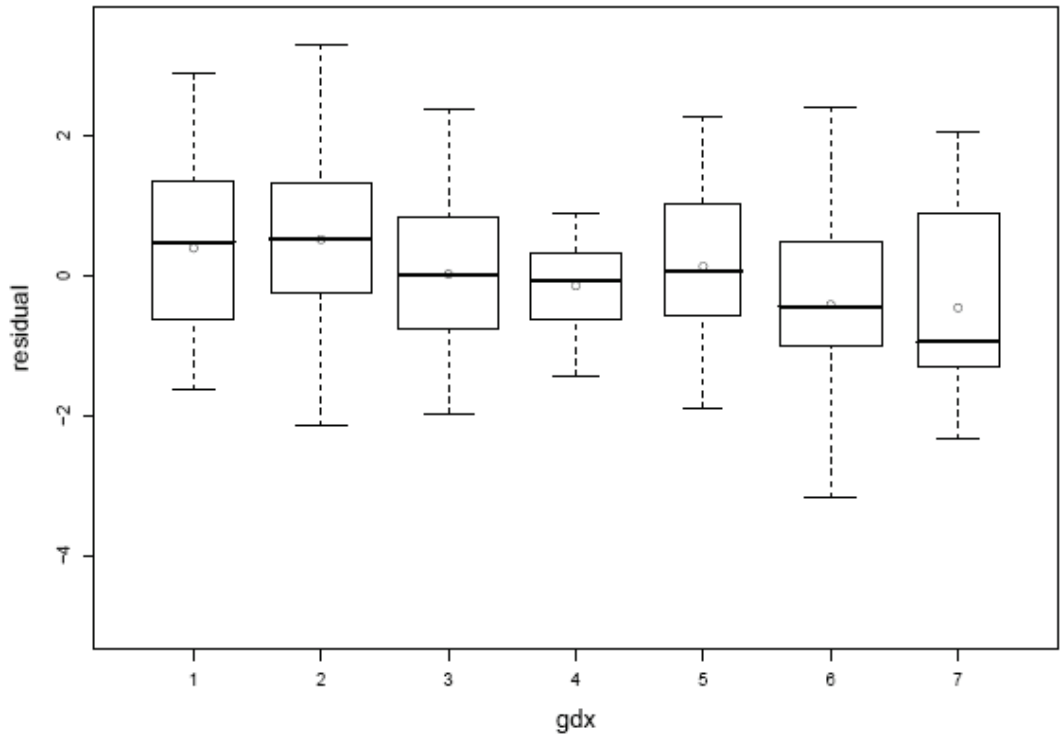

Figure 14. **Box plots for IRT residuals due to item position shift: verbal.**

events allowed one version of this strategy to be evaluated. To do so, item difficulties (transformed *p* values) from samples aggregated across the various scrambled orders were compared to difficulties from the fixed, base ordering.

This comparison simulates the case where items pretested in random positions scattered throughout a section are then administered in fixed positions in operational sections. The hope is that item difficulties based on data collected from a variety of positions would be similar to difficulties based on fixed, central positions.

The further assumption is that center-position difficulties should better predict item performance across the range of operational positions, largely by reducing the distance of possible shifts. That is, items pretested in the center of a section can only shift so far when used operationally, a maximum of one-half the section length. In contrast, items pretested in early or late slots of pretest sections can shift distances equal to nearly the full test section. So the "safest" place to pretest would be center section. Unfortunately, the number of center-section slots is limited under fixed ordering. The intent is then to effectively pretest all items in the equivalent of center position by pretesting each in a scrambled variety of positions.

The (transformed) *p* values aggregated across random scrambles were compared to *p* values from the fixed, base orderings for each item by computing a *t*-test of the null hypothesis of no difference. Since 28 (quantitative) and 30 (verbal) *t*-tests were carried out simultaneously, significance levels were adjusted by the Bonferroni correction (Miller, 1991). Ensuring that the overall Type I error rate is no greater than 0.05 requires a significance level of 0.05/28 = 0.0018 for each quantitative item and 0.05/30 = 0.0017 for each verbal item. Applying these levels to the *t*-statistics shown in Tables 9 and 10 reveals that none of the differences was significant.

The performance of items pretested in multiple positions therefore appears to adequately predict future performance regardless of what fixed position they might take in an operational section. Together with the previous finding that an item needs to be moved forward or backward a distance equivalent to half or more of the section length to induce position effects of a serious magnitude, it is concluded that pretesting items in random locations throughout the test can effectively mitigate position effects, at least under the speededness conditions encountered with this test and this examinee population.

Table 9

Comparison of Scrambled and Base p Values: Quantitative

	Item *p* value			Significance
Item	Scrambled	Base	*t*-statistic	*p* value
Q-1	0.737	0.732	0.27	0.787
Q-2	0.745	0.753	-0.49	0.622
Q-3	0.652	0.659	-0.38	0.705
Q-4	0.779	0.804	-1.55	0.121
Q-5	0.812	0.829	-1.14	0.253
Q-6	0.857	0.894	-2.83	0.005
Q-7	0.223	0.197	1.64	0.101
Q-8	0.717	0.764	-2.75	0.006
Q-9	0.407	0.387	1.03	0.304
Q-10	0.724	0.719	0.27	0.787
Q-11	0.755	0.770	-0.87	0.383
Q-12	0.805	0.824	-1.28	0.201
Q-13	0.738	0.768	-1.81	0.071
Q-14	0.730	0.738	-0.43	0.664
Q-15	0.577	0.600	-1.18	0.240
Q-16	0.634	0.673	-2.10	0.036
Q-17	0.501	0.538	-1.92	0.055
Q-18	0.665	0.700	-1.94	0.052
Q-19	0.615	0.601	0.76	0.446
Q-20	0.558	0.553	0.26	0.797
Q-21	0.743	0.734	0.54	0.592
Q-22	0.517	0.542	-1.30	0.195
Q-23	0.589	0.600	-0.58	0.559
Q-24	0.584	0.574	0.53	0.598
Q-25	0.656	0.628	1.50	0.134
Q-26	0.691	0.657	1.85	0.064
Q-27	0.470	0.424	2.37	0.018
Q-28	0.742	0.700	2.44	0.015

Note. Degree of freedom for the t-tests is equal to 2,636. Positive *t*-statistics suggest the items appeared more difficult in the base than in the scrambled orderings.

Table 10

Comparison of Scrambled and Base p Values: Verbal

	Item *p* value			Significance
Item	Scrambles	Base	*t*-statistic	*p* value
V-1	0.755	0.754	0.11	0.914
V-2	0.837	0.824	0.90	0.368
V-3	0.541	0.560	-1.00	0.317
V-4	0.404	0.431	-1.37	0.170
V-5	0.540	0.561	-1.08	0.280
V-6	0.481	0.498	-0.86	0.389
V-7	0.646	0.671	-1.33	0.182
V-8	0.279	0.294	-0.80	0.425
V-9	0.818	0.784	2.18	0.030
V-10	0.236	0.234	0.09	0.931
V-11	0.804	0.784	1.24	0.216
V-12	0.730	0.744	-0.76	0.446
V-13	0.636	0.644	-0.43	0.670
V-14	0.596	0.569	1.37	0.170
V-15	0.357	0.378	-1.11	0.267
V-16	0.767	0.776	-0.54	0.592
V-17	0.715	0.710	0.27	0.788
V-18	0.326	0.335	-0.46	0.647
V-19	0.788	0.773	0.88	0.382
V-20	0.428	0.430	-0.10	0.923
V-21	0.713	0.687	1.39	0.165
V-22	0.358	0.371	-0.67	0.504
V-23	0.747	0.708	2.23	0.026
V-24	0.521	0.482	1.98	0.047
V-25	0.738	0.744	-0.35	0.730
V-26	0.287	0.264	1.33	0.185
V-27	0.676	0.652	1.27	0.203
V-28	0.129	0.128	0.10	0.917
V-29	0.640	0.611	1.49	0.135
V-30	0.419	0.419	-0.01	0.995

Note. Degree of freedom for the t-tests is equal to 2,552. Positive *t*-statistics suggest the items appeared more difficult in the base than in the scrambled orderings.

Question 2: Are Item Position Effects Evident in Multistage Tests Administered to GRE Examinees?

Answers to this question were based on analyses of the data gathered during the third collection event, which approximated the administration of multistage tests. However, Data Collections 1 and 2 supported these analyses by collectively administering in scrambled orders all of the items that comprised the multistage tests delivered in Collection 3. The intent was to simulate an operational context where items are pretested (and calibrated) in scrambled linear forms and then used operationally in a multistage test. Data Collections 1 and 2 therefore played the pretest role while Collection 3 stood in for the operational multistage test.

Item response theory (IRT) calibration and linking. Because the three data collections were based on very different examinee samples, care had to be taken to ensure that calibrations from the two pretest data collections were on the same proficiency scale. Because the first and second data collections shared no items in common, IRT parameters estimated from each were linked through the operational GRE scores that were available for examinees. Comparison of these scores showed that the July-August (Collection 1) examinees were both less able and less variable than the September-October (Collection 2) cohort. The relationships between means and standard deviations of proficiency estimates in the two examinee groups can be used to directly determine the linear scaling values needed to link both sets of item parameter estimates to the same scale (Lord, 1980).

However, the linking process was in this case complicated by the fact that the pretest items were calibrated with the 2PL model while the operational GRE scores were three-parameter logistic (3PL) proficiency estimates based on the model that currently underlies the CAT GRE. As noted above, the 2PL model was employed throughout this study because it will be used with the multistage test-based rGRE. Work was therefore required to determine whether it was appropriate to rescale the 2PL item parameter estimates from Data Collections 1 and 2 based on the relationships between the 3PL proficiency estimates for the two examinee groups.

The 2PL item parameter estimates from the two data collections were first transformed to a presumably common scale via the 3PL operational proficiency estimates. The scaled item parameters were then used to produce 2PL proficiency estimates for each examinee from their responses to the linear pretest items. These 2PL estimates were then compared to the 3PL proficiency estimates from the operational CAT. Figures 15 and 16 plot the two sets of estimates

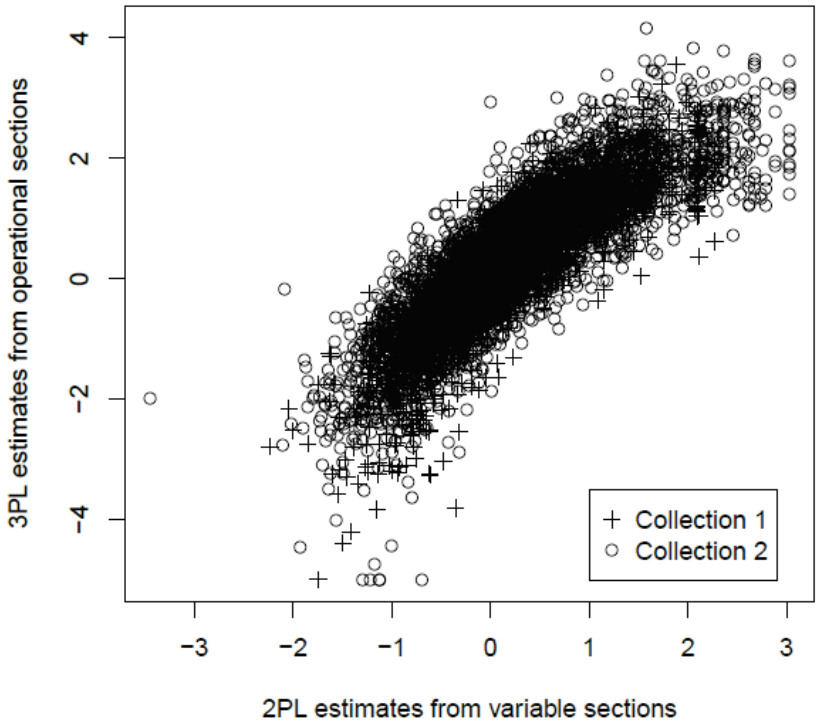

Figure 15. **Scatter plots between 2PL and 3PL estimates for proficiency parameters: quantitative.**

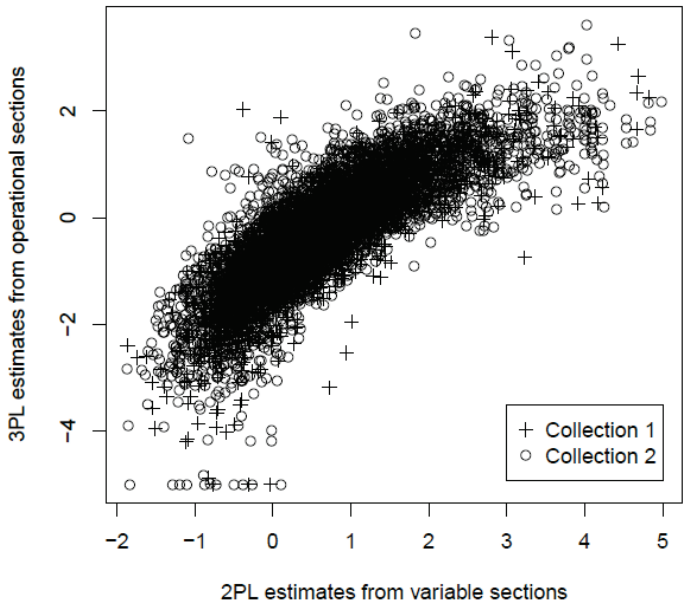

Figure 16. **Scatter plots between 2PL and 3PL estimates for proficiency parameters: verbal.**

against one another and shows an approximately linear relationship over the majority of the proficiency scale. There was little evidence of divergence, or banding, of the relationship across groups, evidence that the 2PL estimates from the two examinee groups may in fact be on a common scale. The correlations between the estimates were 0.85 for quantitative and 0.82 for verbal. Results from the residual analyses of the multistage test data presented below will provide further evidence supporting the conclusion that the item parameter estimates are in fact linked to a common scale.

Residual Analyses

IRT residual analyses served as the primary measure of position effects in the multistage test context. To compute residuals, proficiency estimates were needed for each examinee. These were computed based on the rescaled 2PL item parameters estimated from the scrambled position data from Data Collections 1 and 2.

Unfortunately, proficiencies estimated from the multistage test data will tend to mask item position effects to some degree. This occurrence is because differences in item performance due to position effects will directly influence (or bias) proficiency estimates. For example, suppose that items administered in the last module of the multistage test are indeed operating as more difficult than their random-position item parameter estimates would predict. Then (unexpectedly) high rates of incorrect responses to these items will depress proficiency estimates. These depressed proficiency estimates will, in turn, produce response probabilities that are more in accord with what the pretest item parameter estimates expect, reducing the magnitude of observed residuals. Results presented below will therefore need to be cautiously interpreted.

The same IRT-based residuals described earlier were computed across all of the items administered in the multistage test. For presentation clarity, these were then aggregated within each of the 16 modules. Recall that these modules differ from one another both by their position within the test and by difficulty.

Box plots characterizing the residuals are presented in Figures 17 and 18. Two trends are apparent with both the verbal and quantitative measures. The first is that residuals tend from positive to negative as the multistage test progresses from the first through the last stage. As before, positive residuals indicate that items are operating easier than the IRT parameter estimates predict while negative residuals indicate the opposite. The second, more striking trend

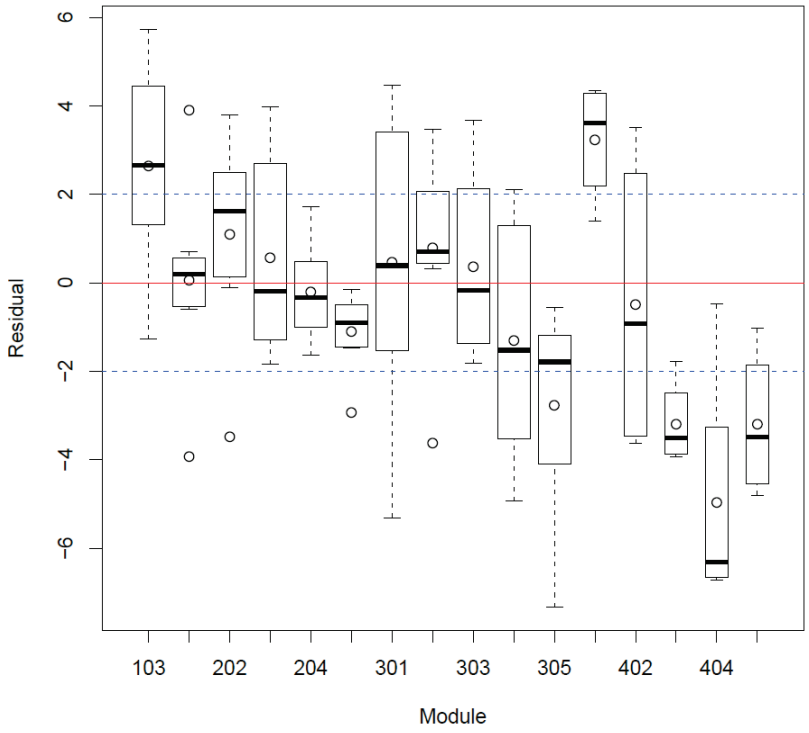

Figure 17. **Multistage test item residuals grouped by module: quantitative.**

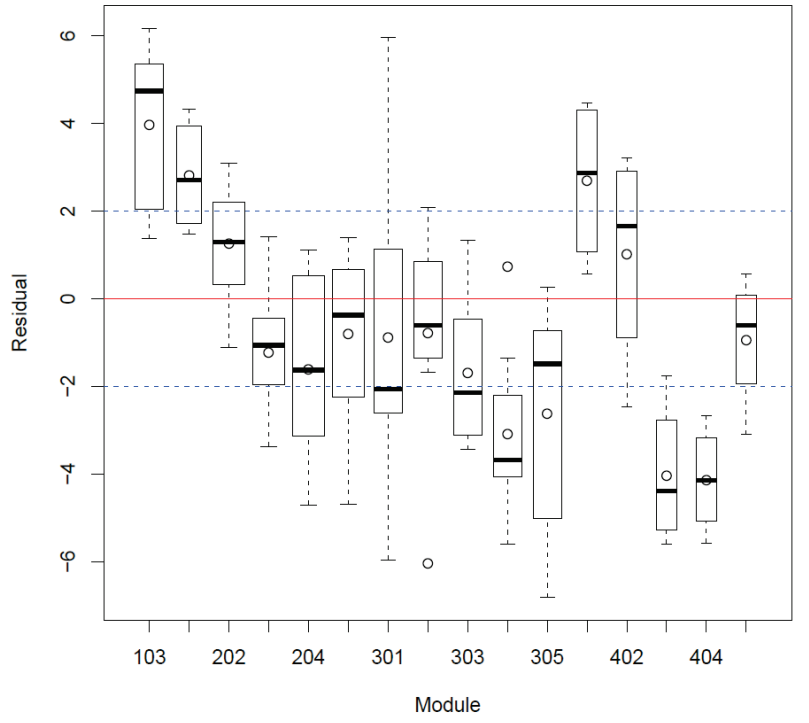

Figure 18. **Multistage test item residuals grouped by module: verbal.**

is embedded within each stage. It shows that modules comprised of easier items are operating as still easier than the IRT parameter estimates predict while modules comprised of difficult items are operating as even more difficult than the IRT parameter estimates predict. This effect seems to grow more pronounced as the test progresses through the final two stages.

For each measure, a regression model that included three predictors was fitted to the residuals. The first predictor indicated the stage (1 to 4) in which an item was administered. The second predictor indicated module difficulty within stage, with the single first stage module coded as 3 while the five modules available in each subsequent stage were coded as 1 (*easiest*) through 5 (*most difficult*). The third predictor indicated the data collection event (1 or 2) from which an item was calibrated. The results of these analyses agreed with the graphical presentation that the effects of stage and module difficulty are pronounced. They also confirmed that the patterns observed in the residuals were not attributable to the different data collection events, strengthening the assertion that the two sets of item parameter estimates were in fact linked to a common scale.

Test Speededness

It was noted above that position effects can be strongly influenced by the circumstances of test administration. Test speededness is a clear example of this influence. The impact of insufficient time is rarely felt equally across all items in a timed section. Instead, examinees generally underestimate the extent to which a section is speeded and spend too much time on earlier items, leaving themselves unduly rushed near the end of the section. This underestimation can cause late-appearing items to look much more difficult than expected.

Speededness is also relative to test difficulty. While a given time limit may be sufficient to comfortably accommodate a set of easy items, a more difficult set of items may appear speeded under the same limit. Similarly, examinees that are more proficient generally find a given test under a given time limit to be less speeded than do less able examinees. Both of these factors are in play and conflated with one another in the multistage tests that were delivered. The adaptive nature of these tests meant that able examinees took a test that was substantially more difficult than the tests delivered to moderate or low-proficiency examinees.

Several analyses were undertaken to evaluate the potential impact of speededness on the results presented in Figures 17 and 18. Because the tests were administered on computer, response times or *latencies* were collected for each response. These could be summed across

items in a module to produce module latencies as well. Because item response latencies are notoriously skewed, they were transformed by logarithms prior to analysis.

The box plots in Figures 19 and 20 characterize the distributions of module latencies across the 16 modules in the multistage tests. The same trends apparent in the residuals are evident in the latencies as well. Latencies appear to diminish with later stages, implying that at least some examinees may be rushing their responses. However, the clearer trend is again within stage, with easier modules exhibiting shorter latencies. Unfortunately, these two trends can be viewed as potentially contradictory. That is, the shorter latencies for later modules are taken as evidence of speededness while the shorter latencies for easier modules are taken as evidence of a lack of speededness.

A simple, supplemental analysis was therefore suggested. This analysis was based on the classic measure of speededness, section completion rates. Although examinees who failed to complete the variable section had been consistently screened from all analyses presented so far, they were necessarily included in the analysis of completion rates.

Examinees were first divided into five roughly equally sized proficiency strata based on their operational GRE scores. Multistage test completion rates were then computed within each of the strata. The results are shown in Figures 21 and 22, where examinee proficiency increases down the page, with the least proficient in the first stratum and the most proficient in the last. The pattern for verbal looks as expected, with completion rates increasing modestly with proficiency. However, the quantitative measure provides a surprise. Here, the least proficient examinees are much more likely to complete the section than are the most proficient examinees.

Table 11 summarizes completion rates for the quantitative multistage tests relative to the linear tests administered during Data Collections 1 and 2. Although differences in overall proficiency across the three examinee cohorts need to be kept in mind, the multistage tests do appear to be more speeded than were the linear (pretest) forms.

7. Discussion and Conclusions

Before discussing the results presented above, it is likely helpful to emphasize the caution that must be taken in extending inferences from the completed studies to the rGRE. First, even though position effects are likely impacted by specific item types and formats, the logistics of the data collection mechanism prohibited administration of tests that included the new item types that will be introduced by the rGRE.

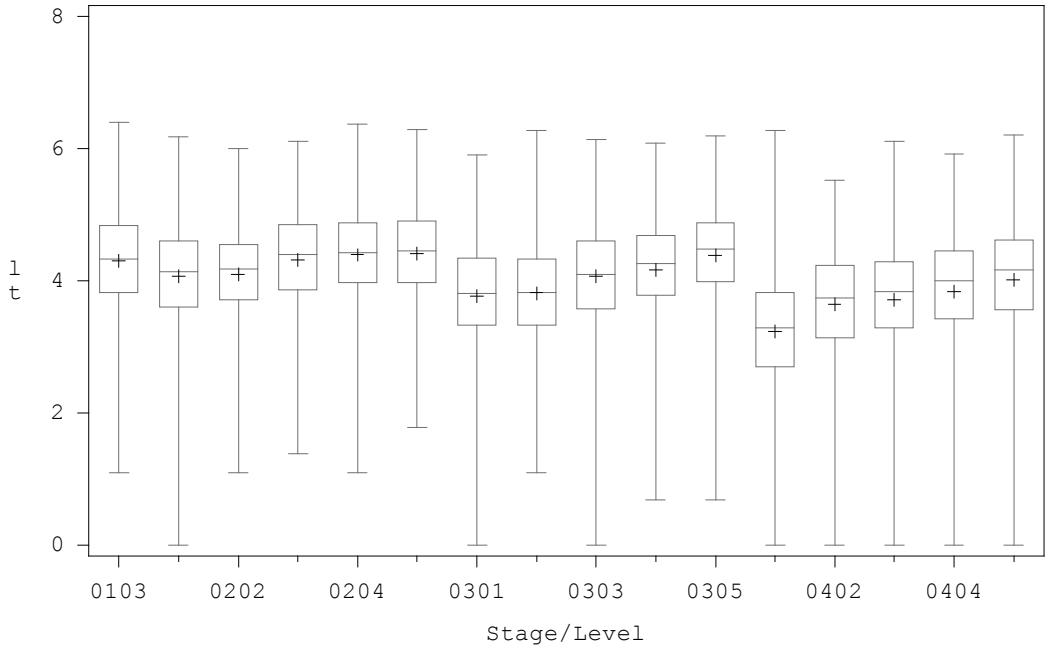

Figure 19. **Response latencies by multistage test module: quantitative.**

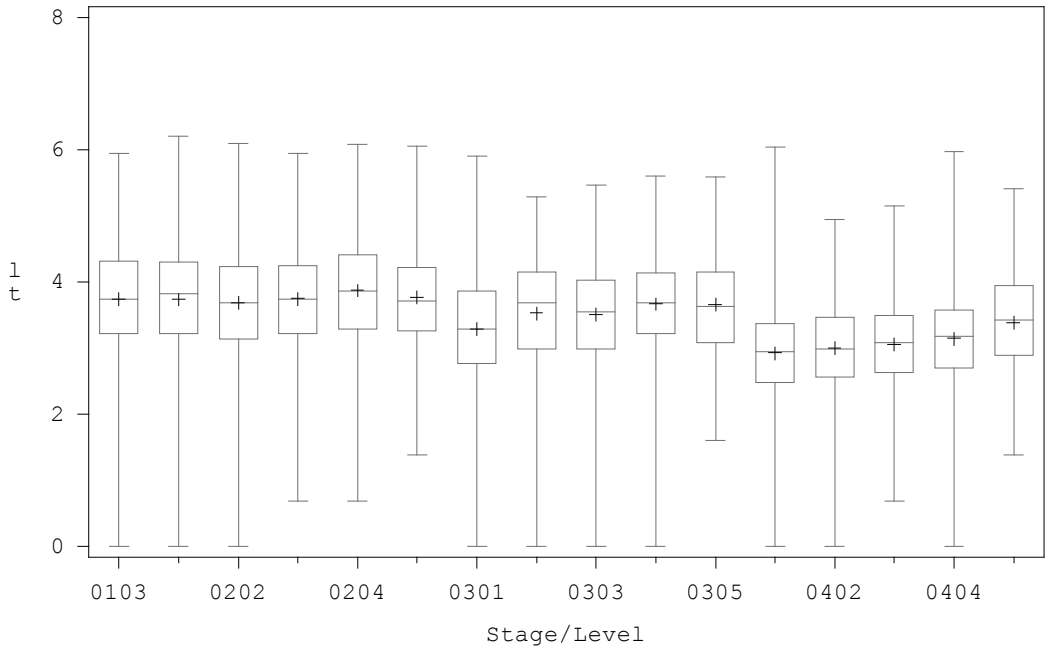

Figure 20. **Response latencies by multistage test module: verbal.**

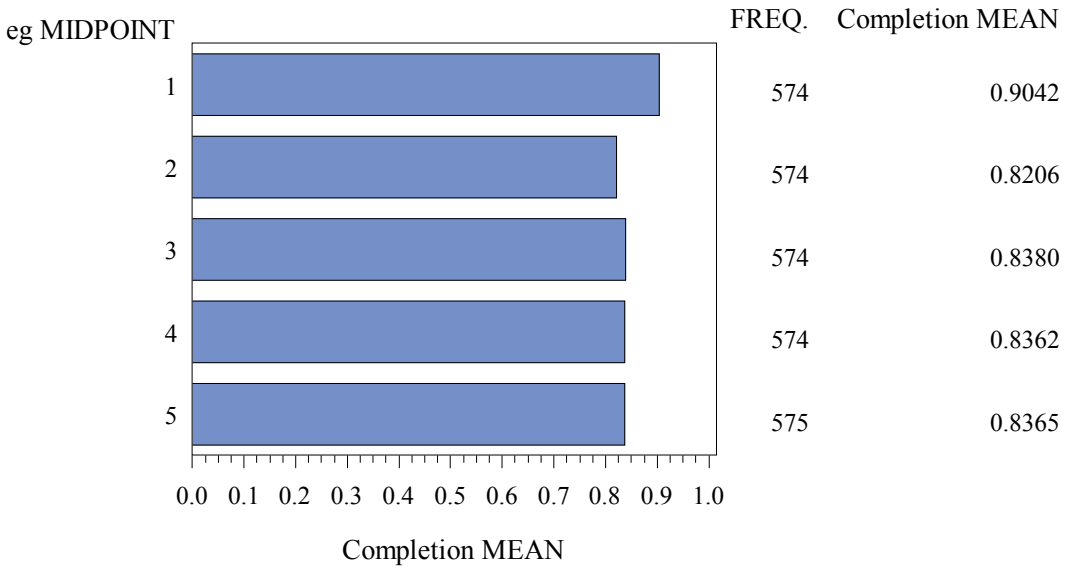

Figure 21. **Multistage test completion rates by proficiency strata: quantitative.**

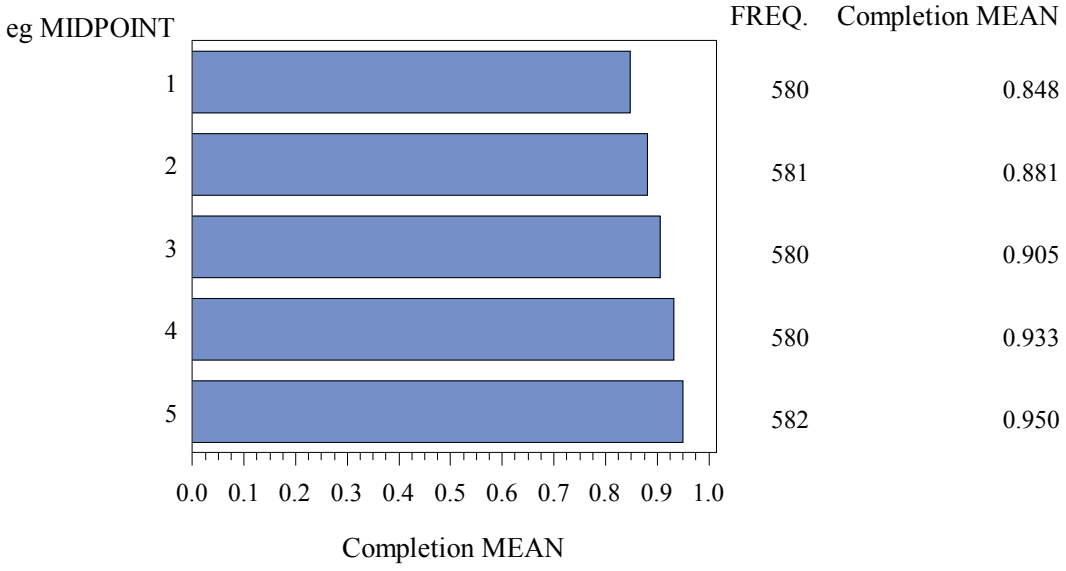

Figure 22. **Multistage test completion rates by proficiency strata: verbal.**

Table 11

Completion Rates for Quantitative Measure by Data Collection Event

Event	Completion rate	Mean operational score[a]
Data collection 1 (July–August, 2008)	91%	.176
Data collection 2 (September–October, 2008)	89%	.394
Data collection 3, MST (March–April, 2008)	85%	-.166

Note. MST = multistage test.

[a] Operational score means are expressed on the internal proficiency metric rather than on the external, reported score scale.

Position effects are also certainly impacted by test speededness, which is itself a complicated function of test length, test difficulty, and time limits. This function becomes still more complicated with adaptive testing, where test difficulty can vary across examinees. Because the variable sections through which study data were collected had to mimic the current GRE in terms of length and time limit, it is unclear what the multistage tests delivered in Data Collection 3 have to say about the multistage tests that the rGRE will deliver.

These substantial caveats aside, the studies offer some compelling results. It is reasonably clear that the linear tests administered during the first data collection exhibit position effects of modest to moderate size. These effects are most evident under the lens of the most sensitive, IRT-based analyses. One possible direction for future analysis is to employ some IRT model that can incorporate item position as a predictor, as well as unique item parameters (e.g., Fischer, 1973; Embretson, 1999). That position effects are pronounced only for the largest shifts in item position—those of half the section length or more—in part validates the strategy of pretesting items in random positions. Because items pretested in random positions approximate the result of pretesting items in central positions, all position shifts will necessarily be limited to less than half of the section length. The results of *t*-tests for the comparison of item *p* values further confirm that the items pretested in a variety of positions throughout the test do not change their performance significantly when compared to their performance in the base ordering. Thus the proposed pretest strategy is able to mitigate position effects to some degree.

Results from the multistage test administrations are both more difficult to interpret and more potentially troubling. However, the most convincing explanation for these results provides

clear guidance to the design of the rGRE. This explanation is that multistage tests, like all adaptive tests, are more subject to speededness than are linear forms of the same length and with the same time limit. Furthermore, speededness is likely to differ across examinees depending on the difficulty of the specific tests each receives.

It is reasonably clear in retrospect that the multistage tests delivered in Data Collection 3 were too speeded, particularly for examinees routed through the most difficult modules. This conclusion is supported by the residual analyses, the module latencies, and the section completion rates. That the impact was particularly pronounced for the quantitative measure is not surprising since difficult quantitative items can have much longer response latencies relative to easier items.

The implications of these studies for the rGRE therefore revolve primarily around the importance of getting time limits set generously enough to minimize speededness, particularly for high-performing examinees. These are lessons the importance of which our years of experience the current CAT GRE has also taught us. We are accordingly engaged in a series of field trials of possible rGRE configurations and time limit combinations, within which careful attention is being paid to speededness.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York, NY: John Wiley.

Beaton, A., & Zwick, R. (1990). *The effect of changes in the national assessment: disentangling the NAEP 1985–86 Reading Anomaly*. Princeton, NJ: ETS.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*, 225–264.

Burke, E., Hartke, D, & Shadow, L. (1989). *Print format effects on ASVAB (Armed Services Vocational Aptitude Battery) test score performance: Literature review*. Brooks AFB, TX: Air Force Human Resources Laboratory.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245–254.

Eignor, D. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT Verbal and Mathematical sections* (Research Report RR-85-10). Princeton, NJ: ETS.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Haladyna, T. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21–25.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Harris, D. (1991, April). *Practical implications of the context effects resulting from the use of scrambled test forms.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kolen, M. & Brennan, R. (1995). *Test equating methods and practices.* New York, NY: Springer.

Kolen, M., & Harris, D. (1990). Comparison of item pre-equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27,* 27–39.

Leary, L., & Dorans, N. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*, 387–413.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Miller, R. G., Jr. (1991). *Simultaneous statistical inference.* New York, NY: Springer-Verlag.

Pomplun, M., & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educations Computing Research*, 30, 243–254.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York, NY: John Wiley and Sons.

Notes

[1] GRE examinees do occasionally surrender and either omit or respond randomly to all or part of a timed section. Such examinees can be spotted by their having a series of very short item response times or by omitting all or part of a section. Interestingly, this behavior is not always limited to the lowest performers and is not always consistent throughout the exam. Small numbers of examinees with moderately high operational scores omit or click quickly through the variable section. Even more interestingly, a similar number of high-performers on the variable section showed the same behavior on the operational section. Although the variable section is disguised to be indistinguishable from an operational section, some examinees believe they can determine whether a section is worth their best effort or not. However, they guess wrong about as often as they guess right.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. xxxxxx